JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 111, D23S28, doi:10.1029/2006JD007598, 2006

Bias-corrected ensemble and probabilistic forecasts of surface ozone over eastern North America during the summer of 2004

J. Wilczak,[1] S. McKeen,[2,3] I. Djalalova,[1,3] G. Grell,[3,4] S. Peckham,[3,4] W. Gong,[5] V. Bouchet,[6] R. Moffet,[6] J. McHenry,[7] J. McQueen,[8] P. Lee,[8] Y. Tang,[9] and G. R. Carmichael[9]

Received 31 May 2006; revised 16 August 2006; accepted 12 October 2006; published 15 December 2006.

[1] A multimodel ensemble air quality forecasting system was created as part of the New England Air Quality Study (NEAQS-2004) during the summer of 2004. Seven different models were used, with their own meteorology, emissions, and chemical mechanisms. In addition, one model was run at two different horizontal grid resolutions, providing a total of eight members for the ensemble. Model forecasts of surface ozone were verified at 342 sites from the EPA's AIRNOW observational network, over a 56 day period in July and August 2004. Because significant biases were found for each of the models, a simple 7-day running mean bias correction technique was implemented. The 7-day bias correction is found to improve the forecast skill of all of the individual models and to work nearly equally well over the entire range of observed ozone values. Also, bias-corrected model skill is found to increase with the length of the bias correction training period, but the increase is gradual, with most of the improvement occurring with only a 1 or 2 day bias correction. Analysis of the ensemble forecasts demonstrates that for a variety of skill measures the ensemble usually has greater skill than each of the individual models, and the ensemble of the bias-corrected models has the highest skill of all. In addition to the higher skill levels, the ensemble also provides potentially useful probabilistic information on the ozone forecasts, which is evaluated using several different techniques.

Citation: Wilczak, J. M., et al. (2006), Bias-corrected ensemble and probabilistic forecasts of surface ozone over eastern North America during the summer of 2004, *J. Geophys. Res.*, *111*, D23S28, doi:10.1029/2006JD007598.

1. Introduction

[2] During the summer of 2004 the International Consortium for Atmospheric Research on Transport and Transformation/New England Air Quality Study (ICARTT/NEAQS-2004) multiagency air quality field program took place in New England. The goals of the NEAQS field study were to understand the meteorological and chemical processes affecting air quality in New England, to evaluate developmental numerical air quality forecast models for this region, and to improve the forecasts from these models.

[3] One method for improving air quality forecasts is through the use of an ensemble forecast system. Increased forecast skill is a well-known advantage of meteorological ensemble forecast systems compared to single deterministic models (*Palmer and Hagedorn* [2006] and *Kalnay* [2003] provide reviews of meteorological ensemble forecasting). In addition to improved skill, ensembles provide quantitative probability information that one cannot get with individual deterministic models. This probability information (e.g., that there is a 70% chance of an ozone violation today) has the potential to significantly improve the value of the predictions that state and local air quality district forecasters routinely provide to the public.

[4] As part of the NEAQS field program, surface ozone forecasts from eight models were collected and displayed in real time. These models include (1) the NOAA/National Weather Service's Eta-CMAQ model, run in a developmental mode in New England during the summer of 2004; (2) the Canadian Hemispheric and Regional Ozone and NO_x System (CHRONOS) model from the Meteorological Service of Canada; (3) A Unified Regional Air-quality Modeling System (AURAMS), also from the Meteorological

[1]Environmental Science Research Laboratory/Physical Sciences Division, National Oceanic and Atmospheric Administration, Boulder, Colorado, USA.
[2]Environmental Science Research Laboratory/Chemical Sciences Division, National Oceanic and Atmospheric Administration, Boulder, Colorado, USA.
[3]Also at Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA.
[4]Environmental Science Research Laboratory/Global Systems Division, National Oceanic and Atmospheric Administration, Boulder, Colorado, USA.
[5]Meteorological Service of Canada, Downsview, Ontario, Canada.
[6]Meteorological Service of Canada, Dorval, Quebec, Canada.
[7]Baron Advanced Meteorological Systems, Raleigh, North Carolina, USA.
[8]Weather Service National Centers for Environmental Prediction/Environmental Modeling Center, National Oceanic and Atmospheric Administration, Camp Springs, Maryland, USA.
[9]Center for Global and Regional Environmental Research, University of Iowa, Iowa City, Iowa, USA.

0148-0227/06/2006JD007598$09.00

Service of Canada; (4) the Weather Research and Forecasting (WRF)–Chem model run at 27 km resolution, provided by NOAA/ESRL/Global Sciences Division; (5) a second WRF-Chem forecast, using a different set of physical parameterization schemes, and run at 12 km resolution; (6) the Baron's Advanced Meteorological Systems, Inc. (BAMS) MM5-MAQSIP model run at 45 km resolution; (7) a 15 km version of the same BAMS model; and (8) the University of Iowa Sulfur Transport and Emissions Model, 2003 (STEM-2K3).

[5] Because of the large number of models that provided ozone forecasts during NEAQS-2004, it was possible to create for the first time a multimodel ozone ensemble consisting of a large number of diverse individual members that use different meteorological models, different chemical emissions inventories, and different chemical mechanisms. Experience from meteorological ensemble forecasts has shown an improved accuracy of multimodel ensembles compared to ensembles generated from a single model but using different initial conditions [*Ziehmann*, 2000; *Stensrud and Yussouf*, 2003], and it seemed likely that the same would apply to ozone forecasts. This expectation has recently been confirmed by *Delle Monache et al.* [2006] who analyzed an ozone forecast ensemble using two meteorological models run at two different resolutions as input to an air quality model. The use of ensembles [*Dabberdt and Miller*, 2000; *Straume*, 2001; *Draxler*, 2002; *Warner et al.*, 2002], and multimodel ensembles [*Galmarini et al.*, 2004a, 2004b] has also recently been proven to be useful for atmospheric dispersion models, which can be more complex than meteorological forecast models, but do not contain the sophisticated chemistry model components present in the NEAQS-2004 air quality models.

[6] An earlier analysis of the NEAQS-2004 data set [*McKeen et al.*, 2005] examined the skill of the ensemble mean ozone forecast using simple statistical measures, and also examined the benefit of correcting the models using the biases calculated over the entire experimental period, which cannot be done for real-time forecasting. In the present analysis we seek to determine the improvement in model forecast skill obtained by applying a simple running mean bias correction technique that can be applied in real-time, and to examine the skill improvement and other benefits of an ensemble model ozone forecast. These two techniques are investigated on their own and in combination. Therefore this analysis extends the work of *McKeen et al.* [2005] by evaluating a bias correction technique that can be implemented in real-time, and also by examining the value of the ensemble using probabilistic skill measures that are commonly used for meteorological forecasts. *Pagowski et al.* [2005] also used the NEAQS data set to investigate the use of a weighted model ensemble, and provided a cursory analysis of ensemble skill using that technique. The present analysis takes a complimentary approach to forming the ensemble, and extends the examination of ensemble model skill within a probabilistic framework.

[7] In section 2 we discuss the models and observations. Section 3 contains a description of the bias correction technique, and an examination of simple skill measures of the models and their ensemble mean. Section 4 contains an analysis of ensemble probability forecasts, and the summary and conclusions are given in section 5.

2. New England Air Quality Study, 2004: Observations and Models

[8] The domains of the models that provided ozone forecasts for the NEAQS-2004 field study are shown in Figure 1. A brief description of each of the models is given below, and a summary of key features of each model are listed in Table 1, including emissions. More detailed descriptions of the models and Web site links for many of the models are given by *McKeen et al.* [2005].

2.1. NOAA/National Weather Service Eta-CMAQ

[9] The ozone forecast capability of this model was in experimental testing mode during the summer of 2004, and was first implemented into operations over the northeast United States in September 2004. It was expanded in operations to cover the Eastern United States in August 2005. The model uses meteorological fields provided by the NWS operational Eta model, which are used to drive the CMAQ photochemical transport model [*Byun and Ching*, 1999]. Conversion of the meteorological data to the CMAQ grids is accomplished with the PREMAQ preprocessor.

2.2. Meteorological Service of Canada CHRONOS

[10] The CHRONOS model has provided operational ozone forecasts for Canada and the northern U. S. since 2001. Meteorological fields are provided by the regional version of the Global Environmental Model (GEM), the Canadian operational weather prediction model. CHRONOS uses the Acid Deposition and Oxidant Model-2 (ADOM-II) chemical mechanism [*Ryerson et al.*, 2001].

2.3. Meteorological Service of Canada AURAMS

[11] The AURAMS model is similar to CHRONOS with regard to its ozone predictions, but was designed to have the additional capability for forecasting regional particulate matter. AURAMS uses the same chemical mechanism, anthropogenic emissions of gaseous precursors, and biogenic emission assignments between vegetative assignments as CHRONOS. The principal differences from CHRONOS are that the AURMAS biogenic emissions are tied to the surface vegetation types specified using BEIS1 [e.g., *Lamb et al.*, 1993], and it runs on a coarser horizontal grid.

2.4. Baron AMS (BAMS) MAQSIP-RT

[12] The Multiscale Air Quality Simulation Platform, Real-Time (MAQSIP-RT) [*McHenry et al.*, 2004] uses the MM5 (version 3.6) mesoscale meteorological model, and an upgraded version of the Carbon Bond 4 (CBM-4) chemical mechanism. Two different horizontal resolution (45 and 15 km) simulations from MAQSIP-RT were available using the identical model physics and chemistry. Two separate 15 km simulations were provided covering different domains. These were combined to generate a single data set used in this analysis. In both cases, the 15 km grids were spawned 6 hours into the 45 km forecast, and used the 6-hour 45 km chemistry and meteorological forecasts as (interpolated) initial conditions.

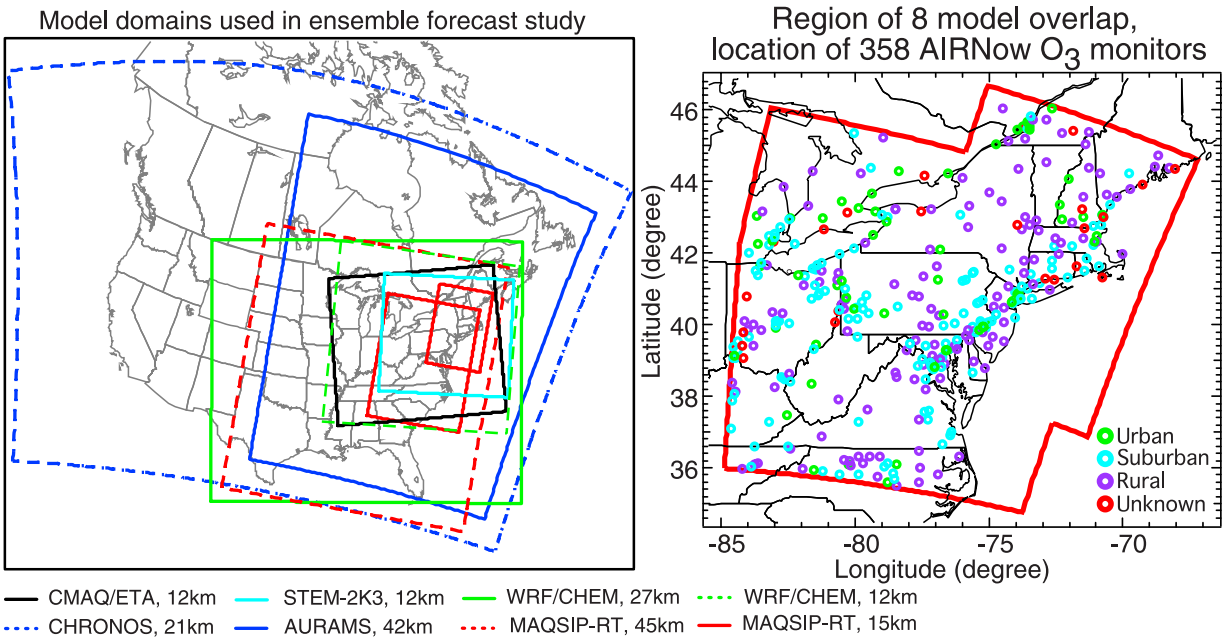

Figure 1. Base maps for the New England Air Quality, 2004, field study: (a) the domains for the eight models that provided forecasts and (b) locations of 342 AIRNOW surface ozone sites in the northeastern United States and southern Canada that are located within the region of common overlap of the eight models.

2.5. University of Iowa STEM-2K3

[13] The University of Iowa STEM-2K3 model [*Tang et al.*, 2003; *Carmichael et al.*, 2003] uses MM5 generated meteorological fields together with the SAPRC-99 gas-phase chemical mechanism [*Carter*, 2000]. The STEM-2K3 model used a nested grid approach, with resolutions of 60, 12, and 4 km. Because the 4 km domain covered only a small geographic area in New England, results from the 12 km grid are used in this analysis.

2.6. NOAA/Forecast Systems Laboratory WRF-Chem

[14] The Weather Research and Forecasting (WRF) chemical model uses the mass coordinate WRF community meteorological model [*Grell et al.*, 2005], combined with gas-phase chemistry based upon the RADM2 chemical mechanism [*Stockwell et al.*, 1990; *Stockwell et al.*, 1995]. Two different model horizontal resolutions were run, at 27 and 12 km. However, since the initialization and physics options were significantly different between the two sets of simulations (Mellor-Yamada-Janjic Eta PBL, 5-class WSM microphysics, Runge-Kutta advection of scalars, Eta initial and boundary conditions for the 27 km runs; YSU PBL, 6-class WSM microphysics, PPM advection of scalars, and RUC initial and boundary conditions for the 12 km runs) the two are considered to be separate models.

[15] All models except for WRF-Chem used meteorological fields calculated offline from their meteorological model components. WRF-Chem is unique among the models in that its chemistry is calculated online at every model time step, with identical model grids used for both the meteorology and chemistry.

[16] Each of the models provided forecasts for at least 28 hours after initialization at 00 UTC (20 LST), except for Eta-CMAQ and BAMS 15 K, both of which were initialized at 06 UTC (02 LST). Therefore daily values of 1-hour and 8-hour maximum ozone are calculated for a 22-hour period beginning at 06 UTC for each model. Also, output from the Eta-CMAQ, CHRONOS, BAMS 15, and BAMS 45 models were each missing 1 day's output during the 56 day experiment. These days with partial model data were included in the statistical analysis.

[17] Hourly averaged ozone values were provided by Sonoma Technology Inc. in real-time from the AIRNow network during the NEAQS-2004 field program for 56 consecutive days from 6 July to 30 August 2004. The observed daily maximum ozone values at each of 342 AIRNow stations are calculated using hourly data over the same 22-hour periods starting at 06 UTC as the models. If more than two hourly averages are missing within a given 22-hour period, a missing value is reported for the maximum ozone on that day.

[18] Although 1-hour and 8-hour maximum ozone concentrations are regulatory quantities of interest, we also investigate model skill on an hourly basis, as variations in skill through the diurnal cycle can provide insights into the physical and chemical causes of model error. The AIRNow observed ozone values are hour averages centered on the half-hour. The Eta-CMAQ model also provided hourly averaged values centered on the half-hour. However, the WRF, STEM, CHRONOS, and AURAMS models provided instantaneous values at the top of each hour. For these models an hourly average value centered on the half-hour was calculated as the average of the two adjacent hourly instantaneous values. The two BAMS simulations provided half-hourly instantaneous values, from which a (1/4, 1/2, 1/4) weighted average of three

Table 1. Key Components and Features of Each of the Eight Air Quality Forecasting Models

Model Name	Met Model	Horizontal Resolution	Vertical Levels (in Lowest 2 km)	Met Initial Conditions/Boundary Conditions	Chemical Mechanism	Anthro Emissions	Biogenic Emissions	Chemical Initial Conditions and Boundary Conditions	Day of Week Anthropogenic Emission Variation
STEM-2K3	MM5	12	9	GFS/GFS	SAPRC-99	NEI-99	IGAC-GEIA	nest within 60 km resolution/MOZART-2	No
BAMS 15	MM5	15	13	Eta/Eta GFS/GFS	CBM-IV	NEI-2001 MOBILE5b	BEIS v3.9 BELD3		Yes
BAMS 45	MM5	45	13	Eta/Eta GFS/GFS	CBM-IV	NEI-2001 MOBILE5b	BEIS v3.9 BELD3		Yes
WRF-Chem12	WRF-ARW	12	16	RUC/RUC	RADM2	NEI-99	BEIS3.11 BELD3	*McKeen et al.* [2002]	No
WRF-Chem27	WRF-ARW	27 km	16	RUC/RUC	RADM2	NEI-99	BEIS3.11 BELD3	*McKeen et al.* [2002]	No
AURAMS	GEM	42 km	14	GEM/GEM	ADOM-II	CEPS-95	BEIS2 Veg.-BEIS1	zero inflow open outflow	Yes
CHRONOS	GEM	21 km	14	GEM/GEM	ADOM-II	CEPS-95	BEIS2 BELD3	zero inflow open outflow	Yes
Eta-CMAQ	Eta	12 km	11	Eta/Eta	CBM-IV	2001 NEI MOBILE6	BEIS3.12 BELD3	CMAQ default profiles	Yes

values was used to obtain an hour average centered on the half-hour.

3. Running Mean Bias Correction and Deterministic Statistics

[19] Weather forecasters have long recognized that post-processing of model predictions will give much improved forecasts of surface variables such as temperature and dewpoint. One of the principal reasons for this is that despite decades of refinement and improvement, meteorological models still contain significant model physics errors. Air quality models, which rely not only on the meteorological model but also on a chemical model, and on highly uncertain emission inventories, are likely to have even greater model errors. A second reason for post-processing of model forecasts is that the point sites used for verification may not be representative of the actual concentrations averaged over an area equal to that of a model grid cell.

[20] Various methods of bias correction are possible. *McKeen et al.* [2005] considered the relative impact of using a mean subtraction bias correction (as used in this study) and a multiplicative, or ''ratio adjustment'' bias correction. The ratio adjustment method provided better skill (using a variety of statistical measures) for those models that had too little ozone variance compared to the observations. However, for models that had a nearly correct variance (as did the mean ensemble) or too large of a variance, the mean subtraction method provided nearly identical RMS errors, but much better Critical Success Index (CSI) values than did the ratio adjustment method. Since in this analysis we focus on the skill of the ensemble, we have chosen to use the mean subtraction method.

[21] Figure 2 shows the mean observed surface ozone for hours 06–28 of the forecast window, averaged over the days of the field program, and over the 342 AIRNOW sites. Predictions for each of the eight air quality models are also shown, using the model gridpoint closest to each observation point for comparison. The diurnal peak of the observed ozone occurs at 21 UTC (17 LST). We note that all of the models have their peak ozone occurring between 1 and 2 1/2 hours earlier than the observed peak. The reason why a bias correction is needed is immediately apparent, with all of the models showing a positive bias in the daily maximum ozone of between 5–25 ppbv.

[22] Also shown in Figure 2 are values for two ozone ensembles. The first is the straight ensemble mean, calculated as the arithmetic average of the eight model forecasts. The second is a bias-corrected ensemble, calculated as follows. First, for each model at each forecast time and at each observation site, the bias is calculated for each of the previous 7 days and then averaged. This bias is then subtracted from today's forecast to obtain a bias-corrected forecast, which is model specific, site specific, and time of day specific. After each individual model is bias corrected, these are averaged to get the bias-corrected ensemble prediction. The dark blue line in Figure 2 shows the ensemble mean of the 8 models, and the red line shows the 7-day bias-corrected ensemble mean, which is virtually identical to the observed value (black line). We note that because the bias correction is linear, the ensemble mean of the bias-corrected models is the same as

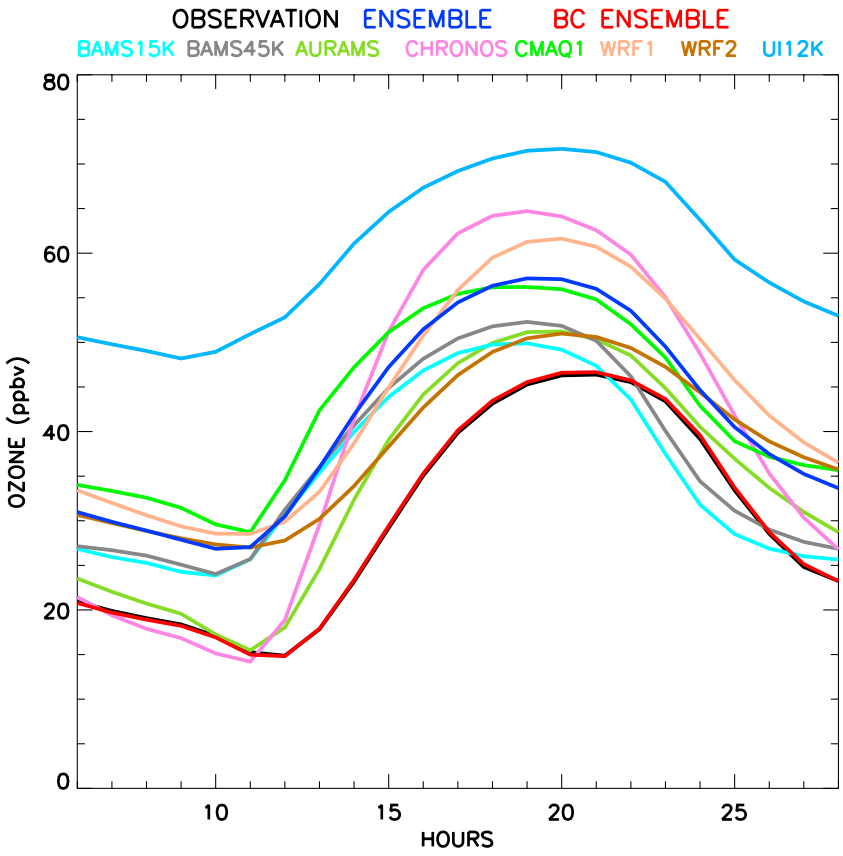

Figure 2. Northeastern U.S. average surface ozone predictions for eight air quality models, their ensemble mean, and the bias-corrected mean, for hours 6–28 of a set of 28-hour forecasts made over 49 days at 342 sites during the summer of 2004, following the color scheme for the various models listed at the top of the figure. Also shown is the observed average surface ozone (black line), which is nearly coincident with the bias-corrected ensemble mean (red line).

calculating the ensemble mean first and then applying a bias correction. However, calculating the bias correction first allows for additional statistics to be calculated for the individual models. Because of the 7-day bias correction, all of the curves shown in Figure 2 are for days 8–56 (49 days total) of the field program.

[23] Histograms of model forecast errors of the daily maximum 1-hour surface ozone are shown in Figure 3,

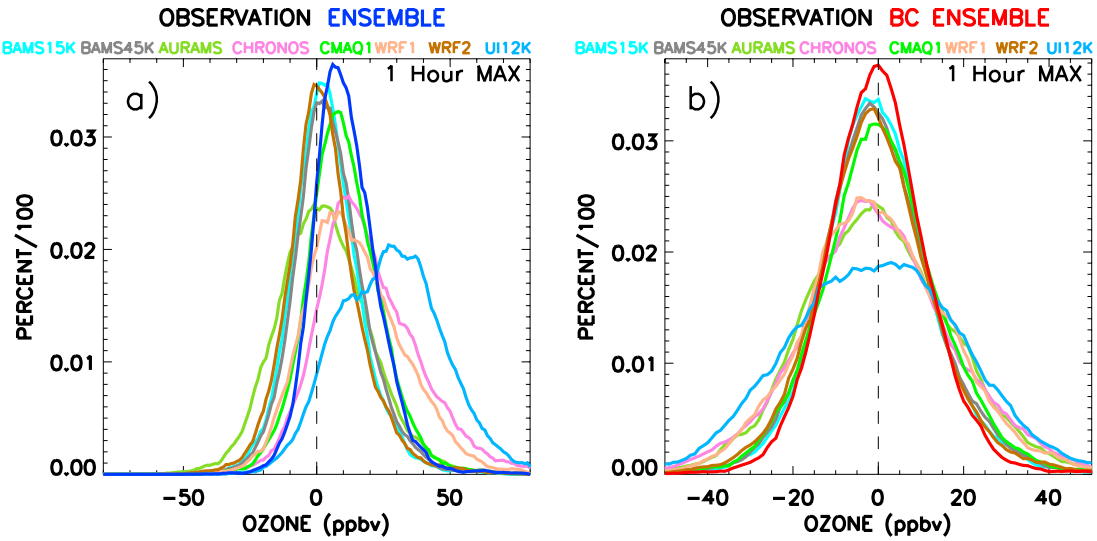

Figure 3. Histograms of model forecast errors (model observation) of the daily maximum 1-hour surface ozone for (a) eight air quality models and their ensemble mean and (b) eight bias-corrected models and their bias-corrected ensemble mean.

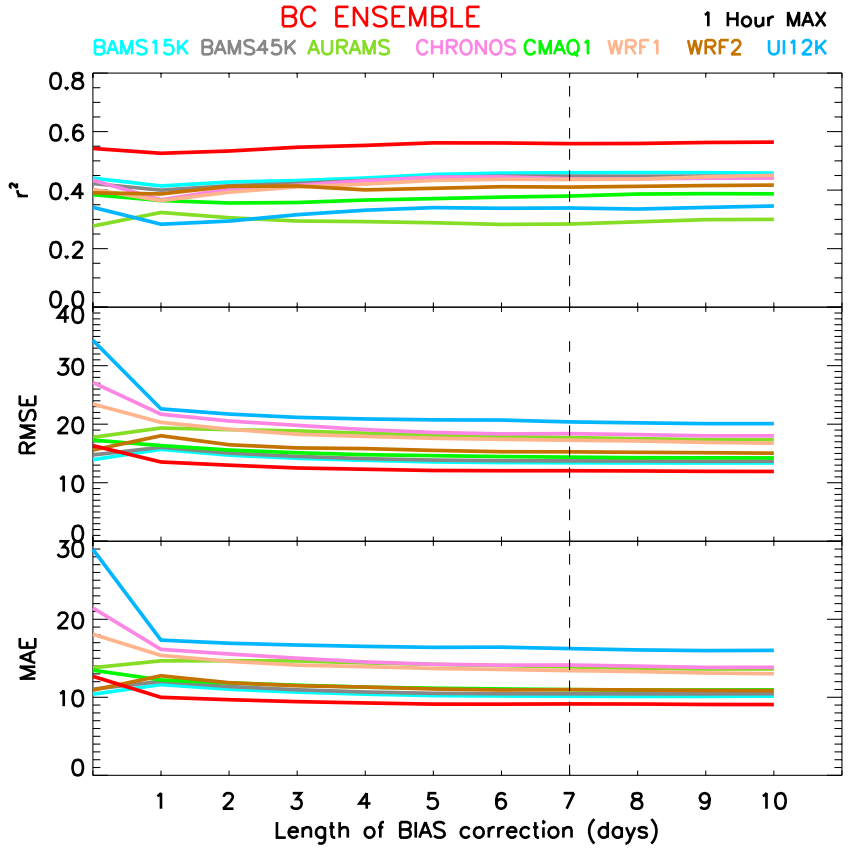

Figure 4. Correlation coefficient squared, RMSE, and MAE for each of the individual bias-corrected models as a function of time for the 1-hour maximum ozone concentration. The bias-corrected ensemble is shown in red.

with and without bias correction. The high bias of the eight models before bias correction is readily apparent in Figure 3a as a shift of the peak of the distribution to toward positive values, as a positively skewed distribution, or both. Application of the bias correction technique to the individual forecast models (Figure 3b) shifts the peak of each error distribution to be close to zero, and reduces the skewness of each model (note the different range of the x axis in the two panels.) Also, the bias-corrected ensemble, shown in red, has a narrower distribution (smaller ozone errors) than any of the individual bias-corrected models with the most predictions with zero error, and the fewest large errors at the tails of the distribution.

[24] The selection of a 7 day period for the calculation of the running mean bias correction was a largely ad hoc choice, based on the assumption that a longer averaging period would be better than a shorter period, up until the point that seasonal effects become important. Figure 4 illustrates the dependence of the square of the correlation coefficient (r^2), RMSE, and MAE [*Wilks*, 1995] on the length of running mean averaging time, for each of the models and their ensemble, from zero days (no bias correction) out to 10 days.

[25] For the 4 models with the largest bias errors (STEM, CHRONOS, WRF1, and Eta-CMAQ) the MAE and RMSE improve for all values of length of the bias correction, even 1 day. For the remaining four models that have smaller bias errors (BAMS 15, BAMS 45, WRF2, and AURAMS), the MAE and RMSE increase for a bias correction length of 1 day, then slowly decrease for longer times. At 7 days the MAE and RMSE for each of the bias-corrected models and the bias-corrected ensemble are smaller than with no bias correction.

[26] The dependence of r^2 on the length of the bias correction is somewhat smaller than MAE and RMSE, and more complex. The significant improvement in r^2 that is obtained from the bias-corrected ensemble compared to even the best performing individual bias-corrected models is readily apparent, and is relatively larger than the improvement found for MAE and RMSE. However, seven of the eight models (the exception being AURAMS) show smaller r^2 values when using a 1-day bias correction compared to no correction. For longer correction lengths between 1 and 10 days, r^2 values for the eight models and the bias-corrected ensemble increase monotonically, with values at 7 days always greater than those with no correction. AURAMS is distinct in that its r^2 value initially increases at 1 day, and then decreases for correction lengths out to 7 days, before increasing again for longer correction lengths. However, even at 7 days, r^2 for AURAMS is larger than it is with no correction.

[27] Next, we compare two simple gross statistical measures of model skill for each of the eight models and ensemble, with and without bias correction, using both

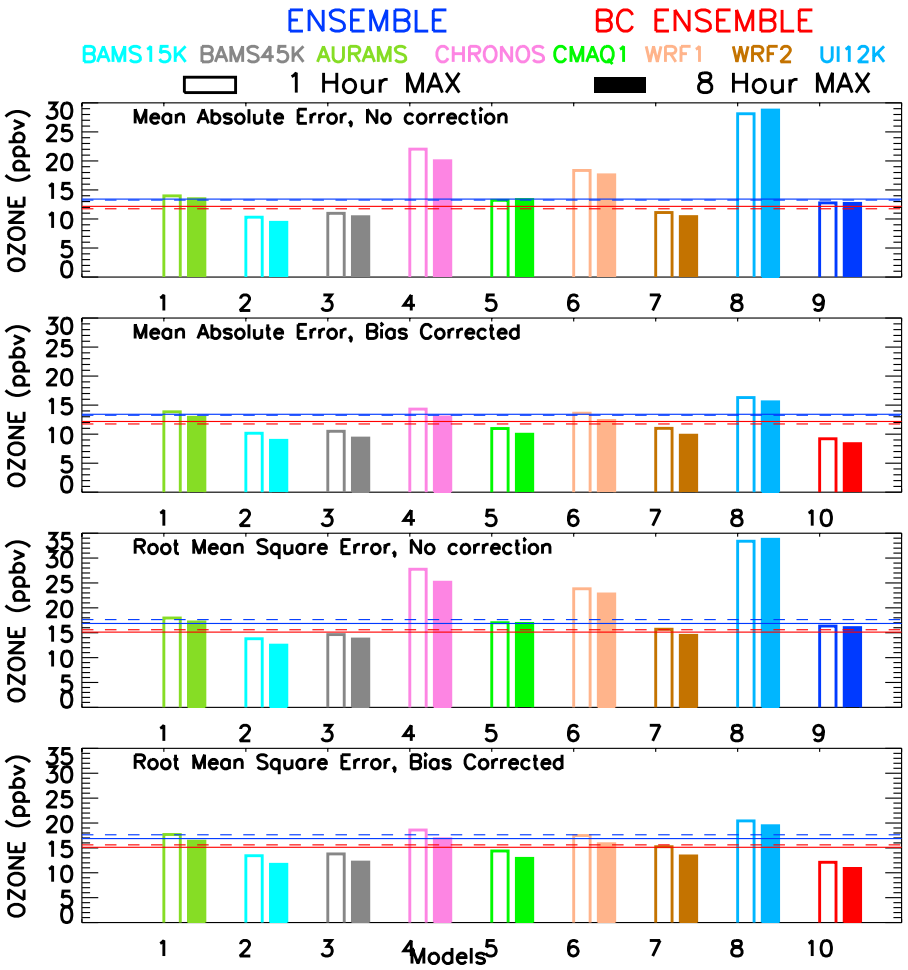

Figure 5. Mean Absolute Error (MAE) (top two panels, with and without bias correction) and the RMSE (bottom two panels, with and without bias correction) for each model. The skill from climatology is shown as a solid horizontal line, persistence is shown as a dashed line, with red for the 1 hour maximum and blue for the 8 hour maximum. The open colored boxes show the 1-hour maximum ozone error statistics, and the solid colored boxes show the 8-hour max ozone error statistics.

1-hour and 8-hour maximum values (Figure 5). Mean Absolute Error (MAE) is shown in the top two panels, and the RMSE is shown in the bottom two panels. For each of these pairs of panels the upper panel contains the uncorrected model statistics, and the lower panel contains the bias-corrected statistics. A pair of columns is shown for each model, with the height of the columns indicative of the size of the error and with the left (open) column showing the 1-hour error and the right (solid) column showing the 8-hour maximum ozone error. Errors using persistence and climatology forecasts are shown as horizontal solid and dashed lines, with persistence being the error using a forecast value equal to the previous days observed 1-hour maximum ozone value and climatology being the error using a forecast value equal to the mean daily 1-hour maximum ozone value observed over the entire field program. The persistence and climatology lines are the same in the top two panels and in the bottom two panels.

[28] From Figure 5, several points are readily apparent. First, the 1-hour and 8-hour maximum ozone error statistics are very similar, with a slightly smaller error for the 8-hour values for almost all of the models. The smaller errors for 8-hour maximum ozone are not surprising, as the longer averaging time helps reduce short timescale meteorological variability (e.g., cloudiness, wind direction shifts, etc.) that will be difficult for a model to reproduce accurately and that can significantly alter the 1-hour peak value. Second, for every individual model, the bias correction technique improves the MAE and RMSE. Third, the skill of the ensemble is very similar to both persistence and climatology, but the bias-corrected ensemble is considerably better than both. The percent reduction of error for the bias-corrected ensemble compared to the uncorrected ensemble is 28% for 1-hour MAE, 34% for 8-hour MAE, 26% for 1-hour RMSE, and 32% for 8-hour RMSE. Finally, one pair of model forecasts were run with identical models, but a factor of 3 difference in horizontal model resolution (BAMS 45 K and BAMS 15 K). Despite the significantly higher horizontal resolution of BAMS 15 K, it has only marginally improved MAE and RMSE statistics.

[29] The bias correction technique so far has been found to improve the gross statistics of all of the models and their ensemble, whether considering all ozone values or daily 1-hour and 8-hour max values. However, because days

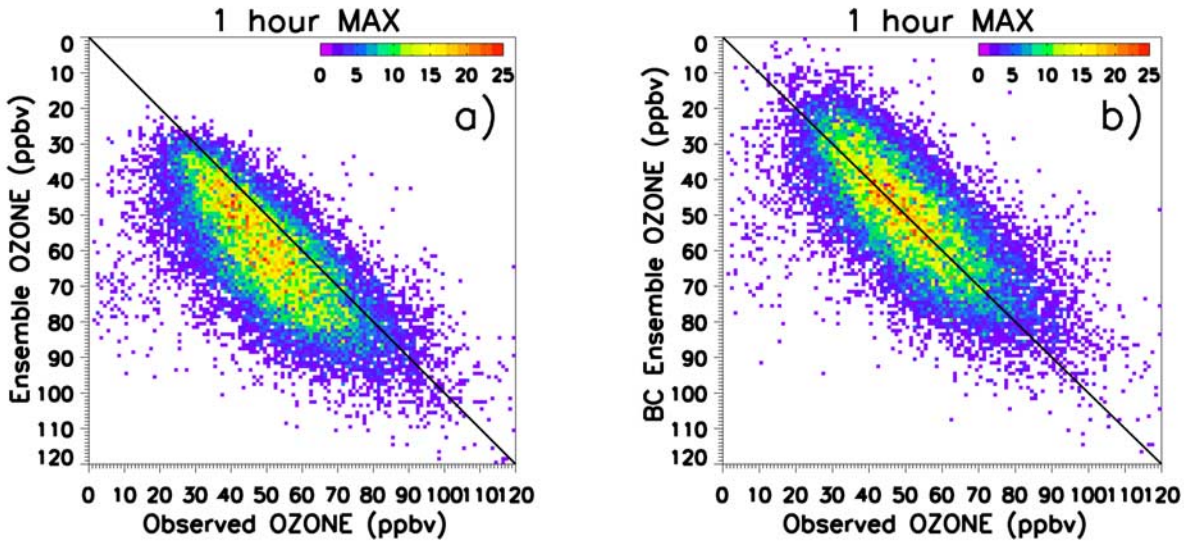

Figure 6. Scatterplot of the ensemble ozone prediction of the daily maximum 1 hour versus the observed ozone for (a) without bias correction and (b) with bias correction. Colors indicate the number of occurrences that fall within 1 ppb bins.

with high ozone are of the greatest concern, it is important to know whether the bias correction technique works equally well on high ozone and moderate or low-ozone days. Frequently this is done by considering the statistics of ozone violations, periods when the ozone exceeds the thresholds of 120 ppbv for the daily 1-hour maximum, and 85 ppbv for the 8-hour maximum. The summer of 2004 in the Northeastern U. S. as was atypical for ozone, however, with few ozone violations, with only 4 out of 19,097 (0.02%) daily 1-hour maximum concentration observations exceeding 120 ppb. Therefore we examine scatterplots of the observed versus ensemble and bias-corrected ensemble predictions of 1-hour maximum ozone (Figure 6), and consider the effect of the bias correction technique at differing values of observed ozone concentrations.

[30] As seen in Figure 6a, the ensemble overpredicts the 1-hour maximum ozone on average for the entire range of observed ozone values, although the magnitude of the overprediction decreases toward the high end of the observed maximum ozone values. The bias correction technique results in near zero residual bias for observed ozone values less than about 85 ppb, but has a tendency to produce too large of a correction for observed ozone values greater than 90 ppb (Figure 6b). However, we note that the number of observations greater than 85 ppb is still relatively small (3.2% of the total number of daily 1-hour max observations), and that it would be useful in the future to test the bias correction technique on a data set with a greater number of high ozone values.

[31] Additional statistical inferences can be drawn from the creation of a contingency table, which is a scatterplot of the ozone data stratified into distinct categories. In our analysis we stratify the model and observed ozone data into 10 ppb increments, which is a compromise between having a wide enough bin to get a statistically meaningful number of points within each bin, but sufficient resolution to delineate trends in the statistics. Figure 7 displays five common statistics of the categorical data [*Wilks*, 1995] as a function of the observed 1-hour maximum ozone concentration (divided into 10 ppbv ozone bins), for both the ensemble and bias-corrected ensemble. These statistics are the Frequency Bias (FB), Percent Correct (PC), False Alarm Ratio (FAR), Probability of Detection (POD), and Critical Success Index (CSI).

[32] The first statistic, the frequency bias, is the number of forecasts divided by the number of observations within each ozone bin, and ideally should be unity. FB for the ensemble is low at small values of observed 1-hour max ozone, and increases monotonically with increasing ozone. For the bias-corrected ensemble FB is near unity for ozone values up to 80 ppb, but then becomes small for higher ozone values. This indicates that the raw ensemble has a smaller bias on the highest-ozone days than on low or moderate ozone days, as seen in Figure 6. The PC is the summation of the diagonal elements of the contingency table divided by the total number of events and also has an ideal value of unity. The PC increases with ozone level, and is larger for the bias-corrected ensemble than for the ensemble, for all values of ozone. The FAR is a measure of the number of false predictions of a given forecast category, and ideally should be zero. The FAR for the raw ensemble is small for low-ozone events, and increases with increasing ozone. In contrast, the FAR for the bias-corrected ensemble is nearly constant, and has lower values than the raw ensemble for ozone values greater than 40 ppb. POD is the fraction of those occasions when the forecast event occurred on which it was also forecast, and has an ideal value of unity. However, we find that the POD is lower for the bias-corrected ensemble for ozone greater than 70 ppb than for the raw ensemble. The better performance of the raw ensemble is due to the high positive bias, which also produces a higher FAR. The CSI is the number of correct

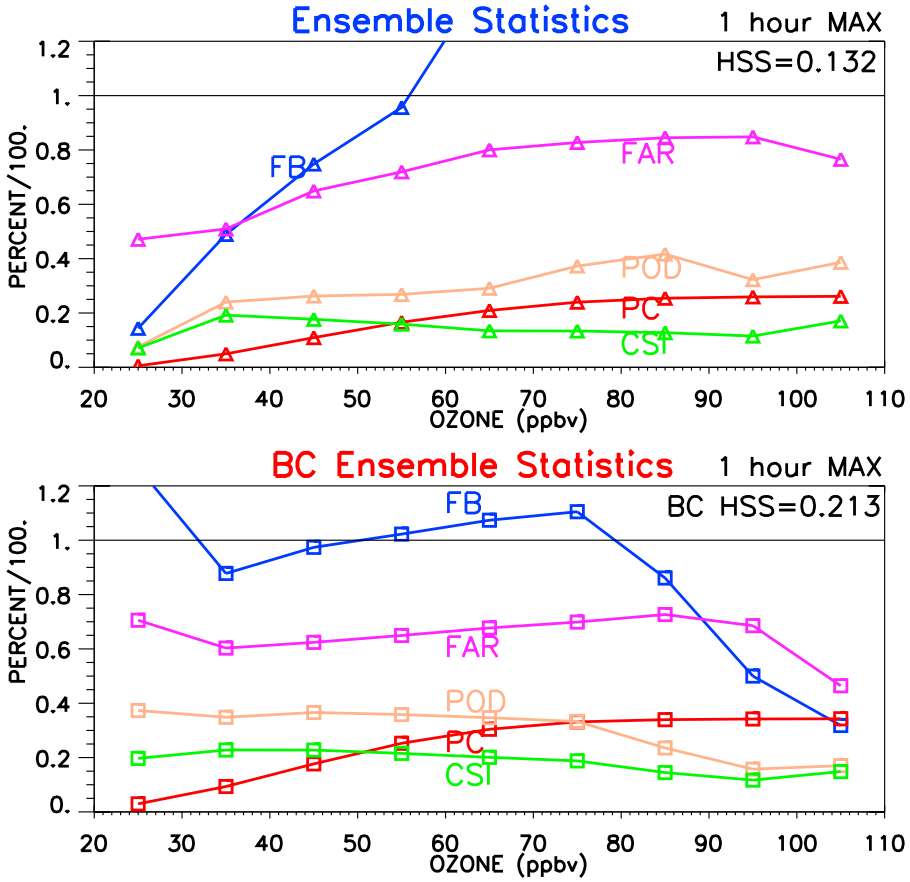

Figure 7. Contingency statistics (frequency bias, percent correct, false alarm rate, probability of detection, critical success index, and Heidke skill score) of the ensemble (top) and the bias-corrected ensemble (bottom).

forecasts divided by the number of cases forecast and/or observed, and ideally should be unity. The CSI for the bias-corrected ensemble is greater than for the raw ensemble for observed ozone values less than 90 ppb, and is essentially identical for ozone greater than 90 ppb. Finally, the Heidke Skill Score, which is the percentage improvement in forecast accuracy compared to random chance, is almost 50% greater for the bias-corrected ensemble than for the raw ensemble.

4. Ensemble Probabilities

[33] In addition to the higher skill shown by the ensemble forecasts compared to individual models, a second advantage of ensembles is that they provide probabilistic information that allows for the uncertainty of a forecast to be expressed quantitatively. We consider a variety of methods for expressing the skill inherent in ensemble forecasts that have been developed and/or applied by the meteorological forecast community, and use them to asses the value of ensemble ozone forecasts.

4.1. Rank Histogram

[34] One standard method for evaluating the behavior of an ensemble is through the rank histogram (see *Hamill* [2001] for an extensive discussion of the rank histogram). A rank histogram is constructed by combining the n-member predictions of the ensemble with the verifying observation into a vector of length n + 1, and then sorting them from highest to lowest (ozone) value. The rank of the verification is then tallied, and a histogram computed by repeating the process over many independent sample points. Ideally, the verification values should have an equal probability of occurring at each position in the vector, indicating that the ensemble member forecasts and the observed state can be considered to be random samples from the same probability distribution.

[35] The rank histogram (Figure 8) of the uncorrected ensemble members for 1-hour maximum ozone shows a strongly sloped distribution associated with the positive biases of the ensemble members, as the observed ozone value is more frequently the lowest member of the vector, and rarely is the highest. In contrast, the bias-corrected rank histogram has almost no slope. Meteorological ensembles often display a U shaped or concave rank histogram, indicating that the ensemble has too little variability and is underdispersive [e.g., *Hamill and Colucci*, 1997; *Stensrud and Yussouf*, 2003]. The flat rank histogram of the bias-corrected ozone ensemble shown in Figure 8 indicates that it contains the proper amount of variability.

4.2. Attributes Diagram

[36] The forecast probability that the 1-hour max ozone will exceed a given threshold can be equated with the

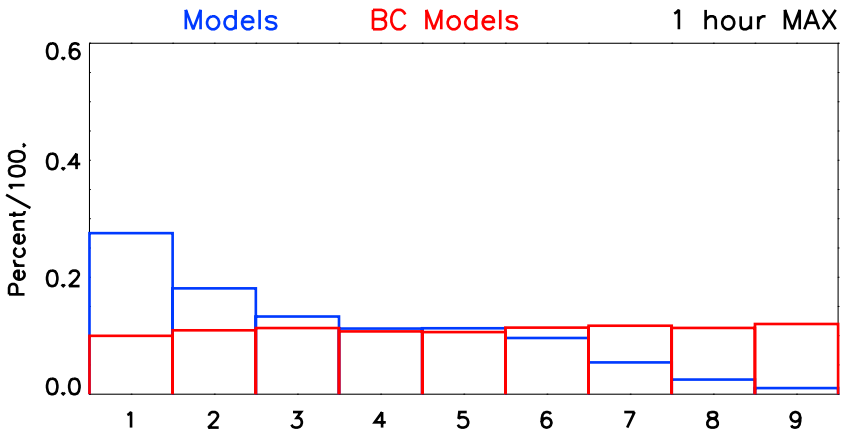

Figure 8. Rank histogram for the ensemble (blue) and bias-corrected ensemble (red).

fraction of the 8 models in the ensemble that exceed that threshold. An attribute diagram [*Wilks*, 1995] depicts the skill of this probability forecast, and is simply a plot of the forecast probability versus the observed frequency of occurrence. The attributes diagram for a threshold of 70 ppb is shown in Figure 9; similar probability curves are found for a wide range of threshold values. Ideally, the reliability curve should follow the 1-1 slope. The ensemble curve shown is blue lies below the ideal line for all forecast probabilities, consistently overpredicting the probability that the ozone would exceed 70 ppb. The bias correction technique reduces by about one-half the tendency for overprediction at this threshold level, bringing most of the forecast probabilities above the no-skill line. Attempts to further improve the reliability of the forecasts were made by using the information from the rank histogram to calibrate the ensemble probabilities using the method of *Hamill and Colucci* [1997, 1998]. This calibration technique did not however provide any significant improvement to the reliability of either the ensemble or bias-corrected ensemble.

4.3. Relative Operating Characteristic (ROC) Curves

[37] Another way to utilize the probability information inherent in an ensemble forecast is through a Relative Operating Characteristic (ROC) diagram [*Swets*, 1973; *Mason*, 1982]. This diagram compares the false alarm rate (false positives) of a set of forecasts versus the hit rate (true positives) for a given threshold, again shown in Figure 10 for a value of 70 ppb. The ideal forecast has no false alarms and a perfect hit rate, which is the upper left corner. A single deterministic model evaluated with a given observational data set provides a single point on the plot (shown as triangles for the raw individual models and squares for the bias-corrected individual models). In contrast, an ensemble forecast plots as a series of points, in our case when we have 1/8 models forecasting ozone >70 ppb, then 2/8, 3/8, etc. Note that the ensemble and bias-corrected curves fall to the left of the individual models, indicating improved skill. Depending on a forecaster's need to avoid false alarms, various ensemble probability values can be selected. An optimal choice (closest to the upper left corner) for the raw

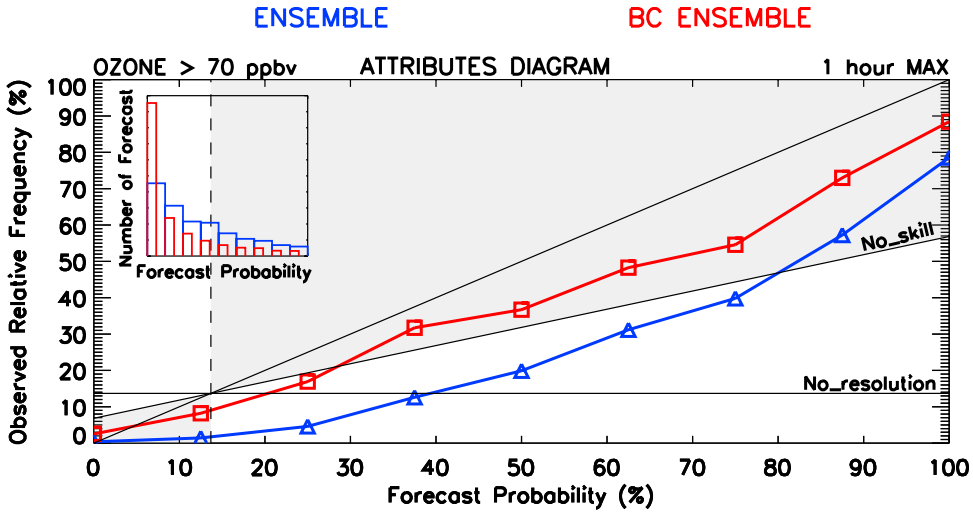

Figure 9. Attributes Diagram for the ensemble (blue line) and bias-corrected ensemble (red line) showing the forecast probability of surface ozone exceeding 70 ppb. Inset diagram shows the frequency of usage of each 10% interval forecast probability category. Horizontal line indicates the climatological frequency of the event in the observed data set (no resolution) and the diagonal dashed line indicates no skill. Ideal probability forecast is indicated by the solid 1:1 slope diagonal line.

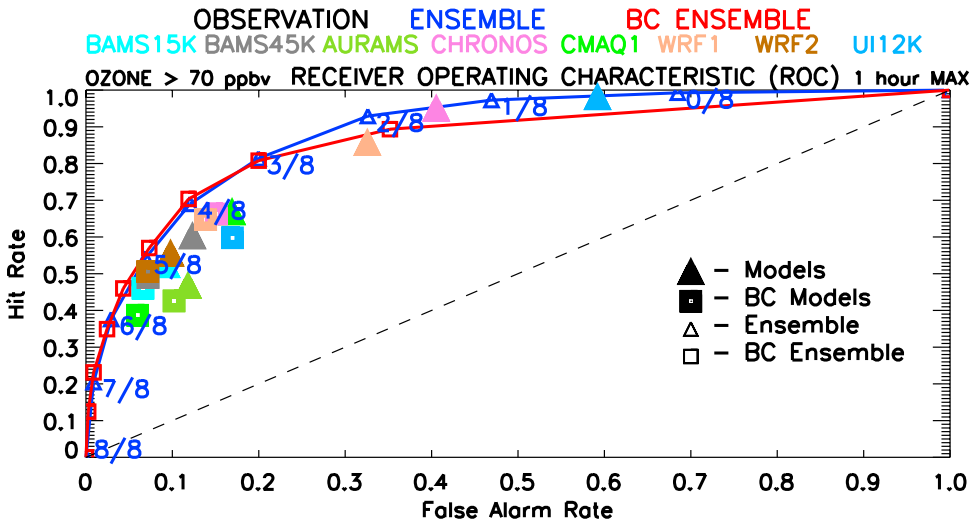

Figure 10. Relative Operating Characteristic (ROC) curves from the ensemble (open triangles, blue) and bias-corrected ensemble (open squares, red) for a surface ozone threshold of 70 ppb for 1-hour maximum ozone. Values in 1/8 increments are shown for the raw ensemble. Solid triangle data points indicate the raw individual models, and solid squares indicate the bias-corrected models, following the color scheme for the various models listed at the top of the figure.

ensemble would be 4/8, but if it was especially important for a forecaster to avoid false alarms, they might choose 5/8 or 6/8 (at the expense of a lower hit rate).

[38] Recently, T. M. Hamill and J. Juras (Common forecast verification metrics can overestimate forecast skill, submitted to *Monthly Weather Review,* 2006) have pointed out that a common way of calculating ROC curves is to assume that the climatology does not vary spatially or temporally among the samples, but that this assumption may produce an overestimation of forecast skill. To circumvent the problem of spatially varying climatology, the ROC values shown in Figure 10 were calculated for each observation location, and these were then averaged. If the ROC values had been calculated relative to a single mean climatology averaged over all of the observation sites, the resulting skill levels would have been ~10% higher. Because of the lack of a reliable long-term ozone climatology, no attempt was made to estimate the contribution of temporal variations in the ozone climatology over the 56 day observation period.

4.4. Spread-Skill Relationship

[39] Ensemble spread is a measure of how well the various ensemble members agree on a given forecast, and is usually taken as the standard deviation of the predicted ozone concentrations. In an ideal ensemble forecast system, one would expect that the spread of the forecasts of the various ensemble members would be related to the skill of the ensemble mean forecast. That is, when the ensemble members disagree on the forecast, the skill of the ensemble mean should be lower. Past studies of ensemble short-range weather forecasts have shown mixed results for a spread-skill relationship. Some have found little correlation between the skill of the forecasts and ensemble spread [*Hamill and Colucci*, 1998; *Stensrud et al.*, 1999; *Hou et al.*, 2001], while others [*Kalnay and Dalcher*, 1987; *Grimit and Mass*, 2002; *Stensrud and Yussouf*, 2003] have found significant correlations.

[40] For the present data set the ensemble spread *SP* is defined as

$$SP(h,d,i) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(C_n(h,d,i) - \tilde{C}(h,d,i)\right)^2} \qquad (1)$$

where C_n is the concentration of ensemble member n, $N = 8$ is the number of ensemble members, $\tilde{C}$ is the predicted ensemble mean concentration, h is the forecast hour, $d = (1, 49)$ is the day of the analysis, and $i = (1, 342)$ is the site location. The skill of the ensemble is defined in terms of the mean absolute error of the ensemble mean prediction,

$$MAE(h,d,i) = \left|C_{obs}(h,d,i) - \tilde{C}(h,d,i)\right| \qquad (2)$$

where C_{obs} is the observed concentration value. The spread-skill relation is the correlation between *SP* and *MAE*.

[41] In calculating the spread-skill relationship of an ensemble forecast system, often the data are averaged before calculating the correlation coefficient. *Hou et al.* [2001] average *MAE* and *SP* over the number of forecast days d, and then calculate a spatial or "pointwise" correlation of the two data vectors of length i at each forecast hour h. Figure 11 displays the spread-skill correlation for the raw ensemble (blue) and bias-corrected ensemble (red) as dashed lines using this temporal averaging approach. The spread-skill correlation for the raw ensemble reaches its maximum value in the afternoon hours just before the ozone peak (see Figure 2), when the boundary layer is well mixed and growing. During the nighttime hours the spread-skill relation drops and even becomes negative. In contrast, the bias-corrected ensemble has a considerably higher spread-skill correlation at all hours, including the late afternoon

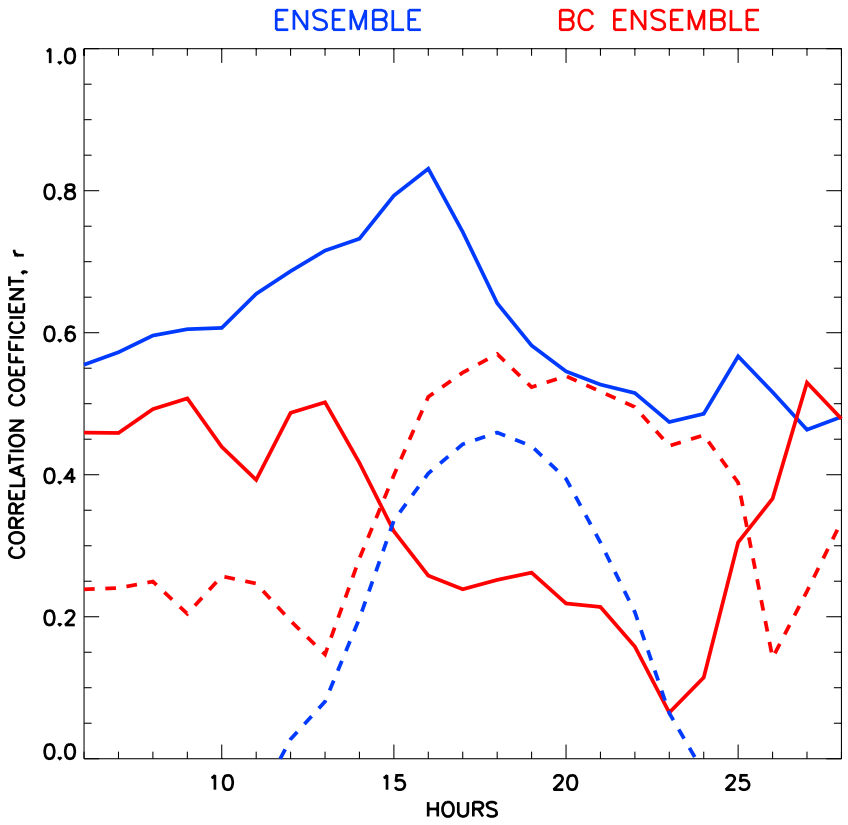

Figure 11. Correlation between standard deviation of the ensemble forecasts (spread) and the absolute error of the ensemble mean (skill), calculated over 49 forecast days and 342 sites at each hour of the forecast cycle. Blue lines are for the raw ensemble, and red lines are for the bias-corrected ensemble. Solid lines indicate that the spread and skill were spatially averaged over all sites and then temporally correlated over the 49 forecast days. Dashed lines indicate that the spread and skill were temporally averaged over the 49 forecast days and then spatially or "pointwise" correlated over the 342 sites.

hours (r = 0.5) when ozone reaches its peak concentrations. We interpret the higher correlation for the bias-corrected ensemble to be the result of the incorporation of information on the spatially varying nature of the bias corrections.

[42] An alternative approach in calculating the spread-skill relation is that of *Grimit and Mass* [2002], who used a spatial averaging approach, in which the spread and skill are averaged over all of the verifying sites i, and then correlated in time though the number of daily forecasts d. We apply this method by averaging the standard deviation of the ensemble forecasts and the absolute error of the ensemble mean from all 342 verification sites, creating two time series of 49 days for each hour of the 22-hour forecast period. These pairs of time series are then correlated and plotted for each hour of the forecast period, and are shown as solid lines in Figure 11 for the ensemble and bias-corrected ensemble. The relative rankings of the raw and bias-corrected ensembles for the pointwise correlation are reversed compared to those found for the temporal correlation method shown as dashed lines. The spread-skill correlation for the raw ensemble is quite high, averaging about 0.65 for all forecast hours. The maximum correlation of 0.83 is found at 1600 UTC (1200 LST), but then drops to approximately 0.55 at the time of peak surface ozone (2100 UTC). In comparison to the raw ensemble, the correlation for the bias-corrected ensemble is significantly lower, averaging about 0.35 and dropping to only about 0.2 in the late afternoon hours at the time of peak ozone concentrations.

[43] The reason for the lower correlation for the bias-corrected ensemble can be seen in Figure 12a, which shows the time series of the daily 1-hour maximum ozone, averaged over the 342 sites, for the 49 days of data available, beginning on 13 July. Also shown in the lower portion of the panel are time series of the ensemble spread and mean absolute error. The MAE and spread both gradually increase through roughly the first 22 days (with some minor fluctuations), decrease significantly around day 24 (5 August), and then increase with time once again. In addition, there are several other minor fluctuations in both spread and skill (at days 3, 12, and 41) that are correlated as well, and the overall correlation coefficient is 0.57. In comparison, the time series of the bias-corrected MAE and ensemble spread (Figure 12b, lower two curves) are relatively constant in time, and have a correlation coefficient of only 0.10. The bias correction technique reduces the overall MAE and spread of the ensemble, but the 7-day filter also reduces the temporal variation of both of these quantities and reduces their correlation.

[44] Finally, we note that if the spread-skill relation is calculated without averaging the data before calculating the correlation, much smaller correlation coefficients are found.

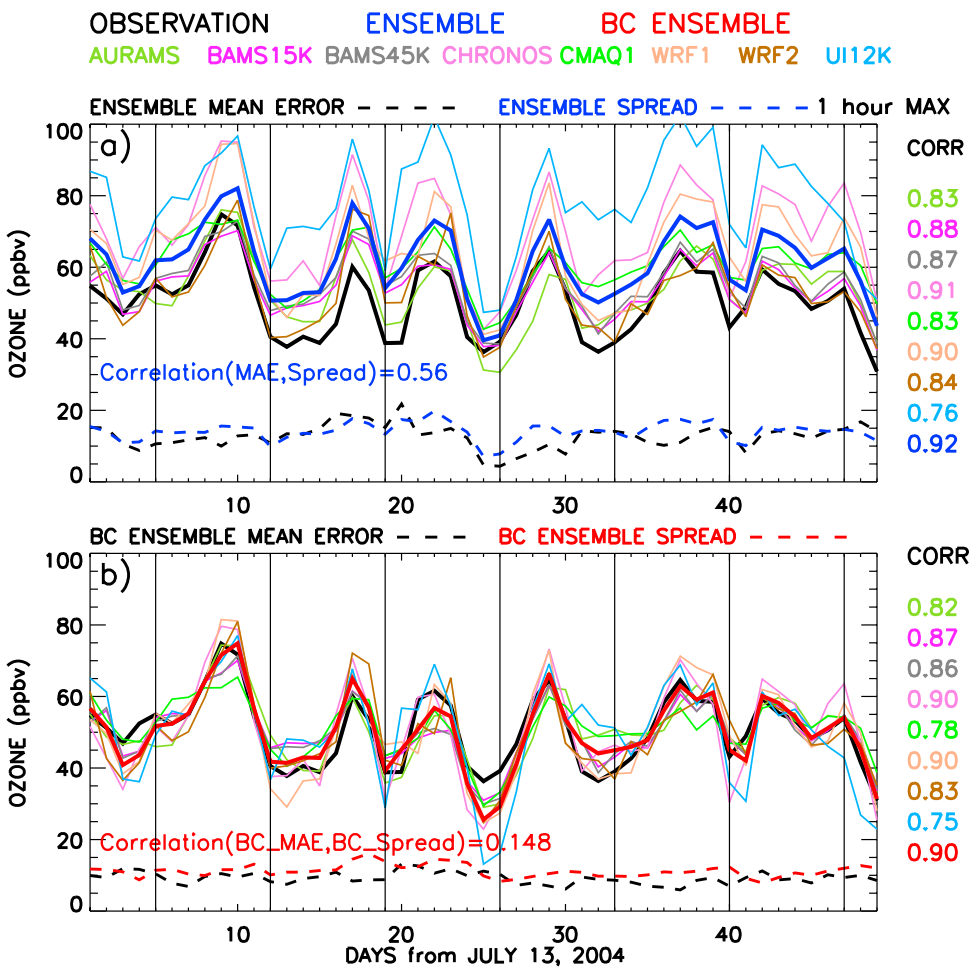

Figure 12. (a) Time series of the daily 1-hour maximum ozone, averaged over the 342 sites, for the 49 days of data available (beginning on 13 July) for each individual model (thin lines), the ensemble mean (thick blue line), and observed values (thick black line). Also shown in the lower portion of Figure 12a are time series of the ensemble spread (blue dashed line) and mean absolute error (black dashed line). Values of the correlation coefficient are shown on the right for each model. (b) Same as in Figure 12a, expect for the bias-corrected individual models and the bias-corrected ensemble (thick red line).

Correlating the spread and skill at each site over the number of forecast days and then averaging over the sites gives diurnal mean values of only 0.25 for the raw ensemble and 0.15 for the bias-corrected ensemble. Calculating the correlations over the sites and then averaging these over the number of forecast days gives diurnal mean values of only 0.15 and 0.13 for the raw and bias-corrected ensembles.

[45] An interesting aspect of the time series of area averaged raw ensemble mean and observed 1-hour maximum ozone shown in Figure 12a (upper curves) is the large amplitude oscillations that occur at approximately weekly intervals. To examine the effect of day-of-the-week variations in emissions on ozone, we also plot in Figure 12a vertical lines for each Saturday. Most of the local minimums of observed ozone occur near weekends, and most of the maximums occur near midweek. However, two out of the seven minima occur on Fridays and one on a Thursday, while one of the local maxima occurs on a Saturday. Also, the temporal correlations of each model and the observed ozone are all quite high, and not significantly different for those models that use day-of-the-week varying emissions (Eta-EMAQ, CHRONOS, AURAMS, BAMS 45, BAMS 15; mean r = 0.86), and those that do not (WRF1, WRF2, STEM-2K; mean r = 0.84). Therefore we conclude that the periodicity apparent in Figure 12a is predominantly due to changes in the synoptic-scale meteorological pattern, which occur on a roughly weekly timescale. We also note the high correlation (r = 0.92) between the time series of the area averaged raw ensemble mean prediction and the observed ozone, as well as for the individual models. The implication is that the models (including their chemical mechanisms) and their ensemble do an excellent job of predicting ozone changes due to the large-scale meteorology. In contrast, if the correlation coefficient is calculated at each site separately and then these coefficients are averaged, the mean correlation is reduced significantly (r = 0.72). This reduction in correlation is due to errors in the spatial variation of the meteorology and emissions in the models. Also we note that the area-averaged bias-corrected ensemble (Figure 12b, upper curves) has almost the same ozone correlation (r =

0.90) than the raw ensemble (r = 0.92). We interpret this to mean that there are no significant trends or slow variations in the area-averaged ozone biases, and that the improvements due to the bias correction technique shown previously are due to its ability to correct for spatially varying biases due to local errors in the meteorology, emissions and site representativeness.

5. Summary and Discussion

[46] The ICARTT/NEAQS-2004 air quality study presented a unique opportunity to create and evaluate a large (eight member) multimodel ozone forecast ensemble. An evaluation of the eight individual models demonstrated that for the summer 2004, during which the eastern U. S. experienced anomalously low ozone concentrations, all of the eight models had significantly positive, diurnally varying biases. Therefore a temporal bias correction technique was applied, where the magnitude of the bias correction was the mean bias for each model calculated at each site and for each hour of the day over the previous 7 days. This technique eliminates the model's biases at each hour of the day, and produces a narrower and more symmetric distribution of ozone errors. Although some individual models have lower MAE and RMSE for 1-hour maximum ozone than the raw ensemble, no individual model or individual bias-corrected model is better than the bias-corrected ensemble. The skill of the bias-corrected ensemble (defined in terms of square of the correlation coefficient [r^2], RMSE, and MAE) increases slowly with correction length. For MAE and RMSE most of the improvement comes after 1 day, while the improvement to r^2 changes more slowly with the length of the bias correction period. The relative improvement of the bias-corrected ensemble over the various bias-corrected individual models is greater for r^2 than for MAE and RMSE. These results differ only slightly if considering 1-hour or 8-hour daily maximum ozone concentrations, with the 8-hour maximum showing slightly better model skill, as the longer averaging time helps reduce short timescale meteorological variability.

[47] The bias correction technique in general produces better categorical skill statistics (frequency bias, percent correct, false alarm rate, probability of detection, and critical success index) for most of the range of observed 1-hour maximum ozone. Exceptions are the FAR for low-ozone events, and the POD of high-ozone events (greater than 85 ppb). However, because of the low ozone values observed throughout the summer of 2004, there are relatively few values higher than 85 ppb which limits the significance of the high-ozone statistics. The bias-corrected ensemble also provides for a higher Heidke skill score than the raw ensemble. Because of the very small number (less than 0.02%) of events during the summer of 2004 where the surface ozone exceeded the daily 1-hour exceedance threshold of 120 ppb, it would be worthwhile to repeat the ensemble categorical analysis for a more typical ozone season.

[48] The use of an ensemble also provides important probabilistic forecast information as depicted in attributes diagrams and ROC curves. The raw ensemble significantly overpredicts forecast probabilities due to the model's positive biases. The bias-corrected ensemble eliminates most of this overprediction of probabilities. ROC curves demonstrate the superior skill of the ensemble and bias-corrected ensemble over each of the individual models or bias-corrected models.

[49] If calculating a spread-skill correlation using a temporal average and correlating in space, the bias correction technique is found to increase the correlation. However, if averaging in space and correlating in time, the bias correction technique reduces the spread-skill correlation by filtering out the temporal variations in skill present in the models. A relatively high spread-skill correlation is found for the spatially averaged raw ensemble (r = 0.64 averaged over all hours) which may be due to the fact that we use a multimodel ensemble where the models employ different physical parameterization schemes.

[50] The high correlation between the spatially averaged ensemble forecasts and spatially averaged observed 1-hour maximum ozone (r = 0.92 for the raw ensemble) indicates that the ensemble (as well as most individual models and their chemical mechanisms) does an excellent job of forecasting the influences of large-scale meteorological variability, and that the large-scale emissions and the chemical mechanisms are generally correct. In contrast, if the correlation coefficient is calculated at each site separately and then these coefficients are averaged, the mean correlation is reduced significantly (r = 0.72). This difference indicates that the greatest improvements to model skill can be achieved through improving spatial variations of the meteorological forecasts as well as improving local emissions variations.

[51] In light of the need to improve the local-scale meteorology, the fact that the high and low-resolution BAMS models (BAMS 15 and BAMS 45) generate almost identical ozone skill levels, indicates that at least in some models, greater horizontal resolution alone may not provide a solution. Since the ensemble provides more skillful forecasts as well as useful probability information, it provides an alternate route for improving air quality forecasts.

[52] **Acknowledgments.** This research was partially supported by Early Starts funding from the NOAA/NWS Office of Science and Technology and the NOAA Office of Oceanic and Atmospheric Research/Weather and Air Quality Program. Credit for program support and management is given to Paula Davidson (NOAA/NWS/OST), Steve Fine (NOAA/OAR), and Jim Meagher (NOAA/ESRL). The authors thank two anonymous reviewers for their constructive comments.

References

Byun, D. W., , and J. K. S. Ching (Eds.) (1999), Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System, *EPA-600/R-99/030*, Off. of Res. and Dev., U. S. Environ. Prot. Agency, Washington, D. C.

Carmichael, G. R., et al. (2003), Regional-scale chemical transport modeling in support of the analysis of observations obtained during the TRACE-P experiment, *J. Geophys. Res.*, *108*(D21), 8823, doi:10.1029/2002JD003117.

Carter, W. (2000), Documentation of the SAPRC-99 chemical mechanism for VOC reactivity assessment, *Final Rep. to Calif. Air Resour. Board Contract 92-329*, Univ. of California, Riverside, Calif., 8 May.

Dabberdt, W. F., and E. Miller (2000), Uncertainty, ensembles and air quality dispersion modeling: Applications and challenges, *Atmos. Environ.*, *34*, 4667–4673.

Delle Monache, L. D., X. X. Deng, Y. M. Zhou, and R. Stull (2006), Ozone ensemble forecasts: 1. A new ensemble design, *J. Geophys. Res.*, *111*, D05307, doi:10.1029/2005JD006310.

Draxler, R. R. (2002), Verification of an ensemble dispersion calculation, *J. Appl. Meteorol.*, *42*, 308–317.

Galmarini, S., et al. (2004a), Ensemble dispersion forecasting, part I: Concept, approach and indicators, *Atmos. Environ.*, *38*(28), 4607–4617.

Galmarini, S., et al. (2004b), Ensemble dispersion forecasting, part II: Application and evaluation, *Atmos. Environ.*, *38*(28), 4619–4632.
Grell, G. A., S. E. Peckham, R. Schmitz, S. A. McKeen, G. Frost, W. Skamarock, and B. Eder (2005), Fully coupled "online" chemistry within the WRF model, *Atmos. Environ.*, *39*, 37,6957–37,6975.
Grimit, E. P., and C. F. Mass (2002), Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest, *Weather Forecasting*, *17*, 192–205.
Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, *129*, 550–560.
Hamill, T. M., and S. J. Colucci, (1997), Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts, *Mon. Weather Rev.*, *126*, 711–724.
Hamill, T. M., and S. J. Colucci (1998), Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts, *Mon. Weather Rev.*, *126*, 711–724.
Hou, D., E. Kalnay, and K. K. Drogemeier (2001), Objective verification of the SAMEX'98 ensemble forecasts, *Mon. Weather Rev.*, *129*, 73–91.
Kalnay, E. (2003), *Atmospheric Modeling, Data Assimilation, and Predictability*, 341 pp., Cambridge Univ. Press, New York.
Kalnay, E., and A. Dalcher (1987), Forecasting forecast skill, *Mon. Weather Rev.*, *115*, 349–356.
Lamb, B., D. Gay, H. Westberg, and T. Pierce (1993), A biogenic hydrocarbon emission inventory for the USA using a simple forest canopy model, *Atmos. Environ., Part A*, *27*, 1673–1690.
Mason, I. (1982), A model for assessment of weather forecasts, *Aust. Meteorol. Mag.*, *30*, 407–430.
McHenry, J. N., W. F. Ryan, N. L. Seaman, C. J. Coates Jr., J. Pudykiewicz, S. Arunachalam, and J. M. Vukovich (2004), A real-time Eulerian photochemical model forecast system, *Bull. Am. Meteorol. Soc.*, *85*, 525–548.
McKeen, S. A., G. Wotawa, D. D. Parrish, J. S. Holloway, M. P. Buhr, G. Hubler, F. C. Fehsenfeld, and J. F. Meagher (2002), Ozone production from Canadian wildfires during June and July of 1995, *J. Geophys. Res.*, *107*(D14), 4192, doi:10.1029/2001JD000697.
McKeen, S., et al. (2005), Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004, *J. Geophys. Res.*, *110*, D21307, doi:10.1029/2005JD005858.
Pagowski, M., et al. (2005), A simple method to improve ensemble-based ozone forecasts, *Geophys. Res. Lett.*, *32*, L07814, doi:10.1029/2004GL022305.
Palmer, T. N., R. Hagedorn (2006), *Predictability of Weather and Climate*, 718 pp., Cambridge Univ. Press, New York.
Ryerson, T., et al. (2001), Observations of ozone formation in power plant plumes and implications for ozone control strategies, *Science*, *292*, 719–723.
Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers (1999), Using ensembles for short-range forecasting, *Mon. Weather Rev.*, *127*, 433–446.
Stensrud, D. J., and N. Yussouf (2003), Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England, *Mon. Weather Rev.*, *131*, 2510–2524.
Stockwell, W. R., P. Middleton, J. S. Chang, and X. Tang (1990), The second generation regional acid deposition model chemical mechanism for regional air quality modeling, *J. Geophys. Res.*, *95*, 16,343–16,367.
Stockwell, W. R., P. Middleton, J. S. Chang, and X. Tang (1995), The effect of acetyl peroxy-peroxy radical reactions on peroxyacetyl nitrate and ozone concentrations, *Atmos. Environ.*, *29*, 1591–1599.
Straume, A. G. (2001), A more extensive investigation of the use of ensemble forecasts for dispersion model evaluation, *J. Appl. Meteorol.*, *40*, 425–445.
Swets, J. A. (1973), The relative operating characteristic in psychology, *Science*, *182*, 990–1000.
Tang, Y., et al. (2003), Impacts of aerosols and clouds on photolysis frequencies and photochemistry during TRACE-P, part II: Three-dimensional study using a regional chemical transport model, *J. Geophys. Res.*, *108*(D21), 8822, doi:10.1029/2002JD003100.
Warner, T. T., R. S. Sheu, J. F. Bowers, R. I. Sykes, G. C. Dodd, and D. S. Henn (2002), Ensemble simulations with coupled atmospheric dynamic and dispersion models: Illustrating uncertainties in dosage simulations, *J. Appl. Meteorol.*, *41*, 488–504.
Wilks, D. S. (1995), *Statistical Methods in The Atmospheric Sciences: An Introduction*, 467 pp., Elsevier, New York.
Ziehmann, C. (2000), Comparison of single-model EPS with a multimodel ensemble consisting of a few operational models, *Tellus, Ser. A*, *52*, 280–299.

V. Bouchet and R. Moffet, Meteorological Service of Canada, 2121 Trans-Canada Highway, Dorval, QC H9P 1J3, Canada.
G. R. Carmichael and Y. Tang, Center for Global and Regional Environmental Research, University of Iowa, 424 IATL, Iowa City, IA 52242-1297, USA.
I. Djalalova and J. Wilczak, Environmental Science Research Laboratory/Physical Sciences Division, National Oceanic and Atmospheric Administration, 325 Broadway, Boulder, CO 80305-3328, USA. (james.m.wilczak@noaa.gov)
W. Gong, Meteorological Service of Canada, 4905 Dufferin Street, Downsview, ON M3H 5T4, Canada.
G. Grell and S. Peckham, Environmental Science Research Laboratory/Global Systems Division, National Oceanic and Atmospheric Administration, 325 Broadway, Boulder, CO 80305-0000, USA.
P. Lee and J. McQueen, Weather Service National Centers for Environmental Prediction/Environmental Modeling Center, National Oceanic and Atmospheric Administration, 5200 Auth Road, Camp Springs, MD 20746, USA.
J. McHenry, Baron Advanced Meteorological Systems, 920 Main Campus Drive, Raleigh, NC 27606, USA.
S. McKeen, Environmental Science Research Laboratory/Chemical Sciences Division, National Oceanic and Atmospheric Administration, 325 Broadway, Boulder, CO 80305-3328, USA.